Data Preprocessing:

To use datasets for further analysis, data preprocessing is necessary. The following key points were considered for pre-processing,

1. Duplicate entries
2. Null values.
3. Column has values in required datatypes such as integer, float, string etc.
4. Few columns has date values, it is necessary to convert and preserve them in datetime format, if dataset is to be used for machine learning algorithms and other visualization.

After cleaning and formatting dataframe is exported to csv file as physical dataset, to be used for further processing.

Pandas us python package used for cleanup and formatting.

Used functions and packages:

- to_datetime() – convert string to datetime
- to_csv() – export pandas dataframe to physical csv file
- readcsv() – load dataset into pandas dataframe
- isna()- check blank values
- isna().sum() – quickly summarize Na values
- duplicated() – check duplicate rows
- duplicated().sum() – quickly sumerise duplicate rows
- Info() – check columns and their datatypes